# EPRINT ARCHIVES – REACHING CRITICAL MASS?

*by Belinda Weaver*

Universities produce varying amounts of research each year. Finding much of this vast output can be problematical. Research findings are scattered across scores of different disciplines, in a range of different scholarly publications or collections of conference papers. Access to this material may be allowed to subscribers only. This limits the availability of research findings and thus limits its usefulness. If people cannot see results, they cannot build on them. Wheels may be unnecessarily reinvented.

This is one of the problems that e-print archives were developed to solve. The aim of an open archive is to make research freely available, in full text, for anyone to use. Some higher education institutions, both abroad, and now here in Australia, are beginning to see the value of archives such as these as a way of showcasing their research in a single, central, searchable space.

*The aim of an open archive is to make research freely available, in full text, for anyone to use.*

What is an e-print? It is simply an electronic version of a paper. The paper may be a book chapter, a conference paper or a journal article. It could be published or unpublished, peer-reviewed or a working paper. This all depends on the collection policy of the individual archive. Some accept theses; others do not. Some will take large data sets so that others can replicate the findings of social scientists and statisticians, or develop them further. E-prints are generally divided into pre-prints (as yet unpublished materials) and post-prints (pieces that have been published, usually in a scholarly journal). Post-prints may be divided up again into peer-reviewed and non-peer-reviewed, depending on the policy of the archive.

The development of e-print archives has been patchy at best, though their future has recently begun to look brighter, as more and more projects get going. The trend towards open archiving of research literature began with the creation of the Los Alamos Physics Preprint archive more than a decade ago, by the trailblazing Paul Ginsparg. The archive began as a pre-print service designed to get round the inevitable delays in scholarly publishing caused by the sometimes lengthy process of writing, submission to a journal, peer review, author's revisions and final publication. Scholars could deposit pre-prints (drafts of new research not yet peer-reviewed) in the archive, thus making the work available for comment, to other researchers, sooner. The Los Alamos archive has been renamed arXiv (**http://www.arXiv.org/**), and has moved under the hosting wing of Cornell University in the US. ArXiv accepts 30,000 papers each year, and many physicists now lodge their papers there instead of submitting them to journals. Accordingly, the service has to some extent re-engineered the process of scholarly communication in physics and the other disciplines, such as mathematics and computing, that the service now embraces.

Other subject-specific archives such as Research Papers in Economics (**http://www.repec.org/**) and CogPrints (**http://cogprints.soton.ac.uk/**), which covers the cognitive and behavioural sciences, have been developed in the past decade, though no service has been as successful as arXiv.

There are several explanations for this, but the most likely is that discipline-specific archives lack effective drivers for development. E-publishing models, based on institutions such as research centres or universities, are now seen as more likely to succeed. Institutions such as these are stable entities, already have established infrastructures, and only need a small reallocation of existing resources to get an archive started. They seem more capable of making the kind of long-term commitment to projects than any attempt to change scholarly publishing will need.

The benefits to an institution of such a service are manifold. An e-print archive can provide a single system to store, and make searchable, research papers produced by staff at a university or research centre. This assists researchers who might be seeking material on a topic, knowing that the university has research strength in that discipline, but who are unaware which academic within the university might have published on the topic. An archive also simplifies archiving for academic staff in different faculties. Instead of academic staff posting and linking to research papers of their own, housed on a school or faculty server, they can deposit papers in a centrally managed system that can promise stable URLs and centralised backup services. This works much better as an academic showcase than an ad hoc system, since, in the latter, there is no clear pathway to such data. This can, in turn, disadvantage academics. The existence of their research online will not assist their visibility if people cannot easily find it.

According to a recent article by Steve Lawrence in *Nature*, 'Online availability of an article may not greatly improve access and impact without efficient and comprehensive search services'. An e-print archive can deliver those types of search facilities by making it possible to locate papers by author, title, keyword, and subject terms. This increases the possibility of a research work's being found, used and cited by others. Articles freely available online are more often cited, according to Lawrence.

➥

Archives also provide an adjunct to the more commonly used research tools of library catalogue and bibliographic databases. Archives may contain collections of materials not available in book or journal form, such as statistical data sets or results of experiments. They may make available in digital form material such as sets of working papers, previously print-only.

Academics interested in inviting comment can notify colleagues of their work-in-progress's availability in an archive. This is much less time-consuming than e-mailing the same piece of work over and over again to different academics. Similarly, academics facing repeat requests for past published papers can point users to archived copies, rather than repeatedly mail out the papers themselves. Academics who wish to use their deposited papers in teaching can direct students to the archive for copies of the work.

Such archives also contribute to the general good. According to the self-archiving FAQ (**http://www.eprints.org/self-faq/**) from eprints.org, 'The purpose of maximizing public access to research findings online is that this in turn maximizes its visibility, usage and impact – which in turn not only maximizes its benefits to researchers and their institution in terms of prestige, prizes, salary, and grant revenue but also maximizes its benefits to research itself (and hence to the society that funds it) in terms of research dissemination, application and growth'.

E-print archives have obviously provided new tools for creating online collections of scholarly work. They speed up the dissemination of research among peers, protect contributors' intellectual property by housing and date-stamping contributions, and guarantee the ongoing availability of electronic research material, something commercial journal publishers still seem loath to do.

Much of the impetus behind the e-print movement has come from librarians who have identified a crisis in scholarly communication. In the existing model, universities are funded by government to produce primary research, much of which goes on to be published in peer-reviewed journals by commercial publishers such as Elsevier, Carfax, Taylor & Francis and others. Academics basically give away their research to academic journal publishers for the glory of being published (which, in turn, advances their careers). Academics also contribute to the system of peer review. They vet the work of other academics, again for no fee. The commercial journal publisher then sells back the research, often at very high cost, to libraries, universities and other research institutions.

What has caused the crisis, if there is one, is the inability of many institutions to maintain key journal subscriptions, especially in the areas of science, technology and medicine where annual journal fees can run into the tens of thousands of dollars. Many libraries have had to cancel subscriptions and plug demand by greater reliance on the inter-library loan system, itself already under strain because of journal price increases and the weak Australian dollar.

Obviously there is something very wrong with this picture, but it is not a situation that can be changed overnight. The system of peer-reviewed publication is still vitally necessary for academic tenure, grant-seeking and promotion, so academics will not happily abandon it for a new one.

However, the development of the Internet and the opportunities it provides for wider and faster dissemination of research at very low costs, have caused many academics and librarians to question the role of academic journal publishers and to seek new ways to get the research out to those who need to use it. Accordingly, initiatives to free the research literature and to safeguard its preservation have begun to spring up.

Professor Stevan Harnad, now at Southampton University, where he helped create the CogPrints archive, was an early advocate of freeing scholarship from commercial constraints. He made a strong case for freeing up the research literature with the publication of his influential article, 'For whom the gate tolls: how and why to free the refereed research literature online through author/institution self archiving, now'.

Initiatives such as the Public Library of Science and BioMed Central aim to challenge, and possibly even replace, existing models of scholarly publishing, as do organisations such as the Scholarly Publishing & Academic Resources Coalition (**http://www.arl.org/sparc/**). Projects such as MIT's D-Space (**http://www.dspace.org/),** OAIster (**http://oaister.umdl.umich.edu/index.html**), the Public Knowledge Project (**http://www.pkp.ubc.ca/),** the Open Archives Initiative (**http://www.openarchives.org/)** and the Budapest Open Archives Initiative (**http://www.soros.org/openaccess/**) are helping to build critical mass for the e-print idea, as are citation analysis tools for archives such as CiteBase (**http://citebase.eprints.org/**). The influential journal, *Nature*, has been running an ongoing forum 'Future e-access to the primary literature' that has tried to draw together all the different points of view on this topic.

So where does this leave Australia? E-print archives are slowly edging over the horizon here, even if their impact within Australia is still quite small. The Group of Eight Universities (ANU, UNSW, Monash, and the Universities of WA, Queensland, Adelaide, Sydney and Melbourne) have agreed to create their own institutional archives. ANU (**http://eprints.anu.edu.au/**) has gone the furthest, launching its archive in 2001. It currently holds several hundred papers. Melbourne has also begun an archive (**http://eprints.unimelb.edu.au/),** and the University of Queensland has started a trial.

The three Group of Eight archives are using the open archive software developed at Southampton University in the UK (**http://www.eprints.org/**). This software is open source and freely available, based as it is on other open source software such as Linux. The software can be customised in-house. The aim of such software is to be compatible with the Open Archives Initiative, thus allowing all metadata from each archive to be harvested, which will

increase the visibility of all work within it. The software offers search and browse facilities for papers in the collection. Users can register with the archive to upload their own work, or can have work uploaded by others. Users can upload different versions of work, for example, an initial pre-print or working paper for comment, revised versions incorporating comments or changes, as well as the final published, peer-reviewed version.

Archive owners can specify the document formats acceptable to them. Since the aim is to make all materials freely available, the most popular formats are PDF, HTML and ASCII text, since these require no expensive software for use. Papers submitted in other formats such as Microsoft Word or PostScript can be converted to PDF.

Subject terms can be allocated to records. Archive owners can choose which thesaurus they wish to use. The e-prints software comes with the Library of Congress Classification but archive owners can use others or develop their own, or forgo a subject list altogether. A logical choice for Australian academic archives would be the Research Fields, Courses and Disciplines, of the Australian Standard Research Classification (ASRC). The ASRC provides a durable, stable thesaurus and is familiar to academics from ARC grant applications. Keywords can also be allocated. This provides a second-string loose subject framework for searchers.

Copyright is the key issue, but even this is not the problem that it might seem, according to eprints's self-archiving FAQ (**http://www.eprints.org/self-faq/**). Many publishers are happy for pre-prints to be deposited and many will allow the post-print as well, provided credit is given to the journal. Getting permission article by article can be time-consuming but many are now streamlining the process with permissions forms. Getting academics on board is another matter. Anyone setting up an archive will need to 'populate' it so that staff can see the usefulness of such a service. Encouraging early adopters is crucial to service take-up. These staff can usually be found in disciplines, such as physics or mathematics, where e-prints are already a familiar part of scholarly research. Once archives reach a certain size, their usefulness becomes more apparent.

After all, they really do work. Roy Tennant, of eScholarship, California Digital Library (**http://escholarship.cdlib.org/**) was quoted in the Free Online Scholarship Newsletter, April 29, 2002:

> 'From the beginning, as we have introduced our repository to our faculty and staff, we have emphasized the point that because they would be depositing their material in an OAI-compliant archive, it would automatically and painlessly be discoverable from various other points around the globe. Luckily, we were right. Within weeks (days?) of opening our doors, we had papers appear in several locations … Formerly, we talked about the possibilities of OAI in the abstract to our faculty. Now we can demonstrate it in reality. That, as you might imagine, is a powerful thing'.

***References***

(2002) *Create change: a resource for faculty and librarian action to reclaim scholarly communication*, Initiative supported by the Association of Research Libraries, the Association of College and Research Libraries, and the Scholarly Publishing & Academic Resources Coalition. **http://www.arl.org/create/**

Crow, Raym (2002) *The Case for Institutional Repositories: A SPARC Position Paper*, Washington, Scholarly Publishing & Academic Resources Coalition. **http://www.arl.org/sparc/IR/ir.html**

Digital Library Federation (2000) *Minimum criteria for an archival repository of digital scholarly journals*, Version 1.2, May 15. **http://www.diglib.org/preserve/criteria.htm**

Frankel, Mark S. (2002) *Seizing the moment: scientists' authorship rights in the digital age* - a report of a study by the American Association for the Advancement of Science, July. **http://www.aaas.org/spp/sfrl/projects/epub/epub.htm**

Gutteridge, Christopher and Harnad, Stevan (2002) *Applications, Potential Problems and a Suggested Policy for Institutional E-Print Archives*. Southampton, Department of Electronics and Computer Science, University of Southampton. **http://eprints.ecs.soton.ac.uk/archive/00006768/**

Harnad, Stevan (2001) *For whom the gate tolls: how and why to free the refereed research literature online through author/institution self archiving, now*. (CogPrints) **http://www.ecs.soton.ac.uk/~harnad/Tp/resolution.htm**

Lawrence, Steve. (2001a) 'Online or Invisible?', *Nature*, 411 (6837): 521. **http://www.neci.nec.com/~lawrence/papers/online-nature01/**

Lawrence, Steve (2001b) 'Free online availability substantially increases a paper's impact', *Nature Web Debates*. **http://www.nature.com/nature/debates/e-access/**

Nixon, William (2002) 'The evolution of an institutional e-prints archive at the University of Glasgow', *Ariadne*, Issue 32, June-July. **http://www.ariadne.ac.uk/issue32/eprint-archives/**

Peek, Robin (2000) E-Prints Are Gaining Momentum : There are many tricky challenges in an Internet-speed world, *Information Today*, v.17, no.9, October. **http://www.infotoday.com/it/oct00/peek.htm**

Pinfield, Stephen, Gardner, Mike and MacColl, John (2002) 'Setting up an institutional e-print archive', *Ariadne*, Issue 31, March-April. **http://www.ariadne.ac.uk/issue31/eprint-archives/**

Pinfield, Stephen (2001) 'How Do Physicists Use an E-Print Archive? Implications for Institutional E-Print

➥

Services', *D-Lib Magazine*, December, v.7, no.12. **http://www.dlib.org/dlib/december01/pinfield/12pinfield.html**

Scholarly Publishing & Academic Resources Coalition (2002) *Gaining Independence: A Manual for planning the launch of a nonprofit electronic publishing venture* **http://www.arl.org/sparc/GI/**

Suleman, Hussein and Fox, Edward A. (2001) A Framework for Building Open Digital Libraries *D-Lib Magazine*, December, v.7 , no.12. **http://www.dlib.org/dlib/december01/suleman/12suleman.html**

Young, Jeffrey R. (2002) '"Superarchives" could hold all scholarly output', *Chronicle of Higher Education*, v.48, no.43.

*Belinda Weaver is the E-print Archive Co-ordinator at the University of Queensland Library.*

NET NOTE

## AUSTRALIAN RESEARCH INTERNET SEARCH TOOL

Research Finder, at **http://rf.panopticsearch.com/search/search.cgi?collection=research**, is an Internet search tool which enables discovery of Australia's researchers, research capability and emerging technologies.

Using P@NOPTIC software (**http://www.panopticsearch.com**), Research Finder allows users to search specifically for Australian research and provides comprehensive coverage of relevant Web sites. Panoptic is a high performance enterprise search engine developed by the CSIRO and the ANU in Canberra. It is a network device (like a fileserver or printserver) which can be installed on an Ethernet network and can be administered via a Web interface. Panoptic includes open source filters for extracting text from common non-HTML formats, such as Word, PDF, PowerPoint, Excel, PostScript and RTF.

The Web sites currently covered by Research Finder include: cooperative research centres (CRCs); the CSIRO; universities; medical research institutes; R&D corporations; technology transfer organizations; and relevant federal government departments and agencies.

Research Finder allows users to carry out free text searches of the Research Finder Index. Searches can be refined by research field, location (i.e. state) and organization type, using metadata generated centrally.


***Figure 1: Research Finder Search Screen***

NET NOTE

## UPDATED VERSION OF CROSSREF

CrossRef (**http://www.crossref.org**) was established by scholarly publishers as an independent, not for profit body in 2000. CrossRef software represents the first full scale implementation of the DOI (Digital Object Identifier) system. It enables researchers to navigate online journals using DOI-based citation links. Version 2.0 of its metadata database resolution services uses a scalable architecture and the latest in Web technology, resulting in more resolved queries (an overall matching rate of 25%), and expanded reference content. The new version also supports the new CrossRef XML Deposit Schema, which provides a more robust vocabulary for journal metadata and conference proceedings. Members have Web access to statistics detailing their deposit and query activity.

CrossRef membership comprises 152 publishers whose content represents over 6,400 journals and almost 5 million article records. CrossRef has recently signed an agreement with Fretwell Downing for the integration of CrossRef into their Z Portal product. Project Muse, which offers subscription access to the full text of over 200 scholarly journals in the humanities and social sciences, will participate in the CrossRef linking program in 2003.